DOCUMENT RESUME

ED 342 806 TM 017 981

AUTHOR Thompson, Bruce
TITLE The Use of Statistical Significance Tests in Research: Some Criticisms and Alternatives.
PUB DATE Apr 92
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Effect Size; *Estimation (Mathematics); Evaluation Methods; *Regression (Statistics); Research Methodology; *Research Problems; Sample Size; *Sampling; *Statistical Significance
IDENTIFIERS Bootstrap Methods; Homogeneity of Variance; *Research Replication

ABSTRACT

Three criticisms of overreliance on results from statistical significance tests are noted. It is suggested that: (1) statistical significance tests are often tautological; (2) some uses can involve comparisons that are not completely sensible; and (3) using statistical significance tests to evaluate both methodological assumptions (e.g., the homogeneity of variance or of regression assumptions) and substantive hypotheses creates inescapable dilemmas. Three strategies for augmenting statistical significance testing are elaborated. First, a review of effect sizes is presented. Second, a method for evaluating statistical significance in a sample size context is discussed. Finally, strategies for empirically evaluating whether results will replicate are reviewed, with an emphasis on explaining one computer-intensive resampling strategy (the bootstrap method). It is inconsistent to use sample results to estimate population values, but to be unwilling to consult the sample to estimate the variability and shape of samples drawn from the population. Three tables and one figure illustrate the discussion, and a 79-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

read.wp1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
BRUCE THOMPSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED342806

The Use of Statistical Significance Tests in Research: Some Criticisms and Alternatives

Bruce Thompson

Texas A&M University

and

Baylor College of Medicine

Paper presented at the annual meeting (session #25.35) of the American Educational Research Association, San Francisco, April 22, 1992.

TM017981

BEST COPY AVAILABLE

ABSTRACT

Three criticisms of overreliance on results from statistical significance tests are noted. It is suggested (a) that statistical significance tests are often tautological; (b) that some uses can involve comparisons that are not completely sensible; and (c) that using statistical significance tests to evaluate both methodological assumptions (e.g., the homogeneity of variance or of regression assumptions) and substantive hypotheses creates inescapable dilemmas. Three strategies for augmenting statistical significance testing are elaborated. First, a review of effect sizes is presented. Second, a method for evaluating statistical significance in a sample size context is discussed. Finally, strategies for empirically evaluating whether results will replicate are reviewed, with an emphasis on explaining one computer-intensive resampling strategy that is often called the bootstrap. It is inconsistent to use sample results to estimate population values, but to be unwilling to consult the sample to estimate the variability and shape of samples drawn from the population.

The use of statistical significance testing as part of the interpretation of empirical research results has historically generated considerable debate (Carver, 1978; Huberty, 1987; Morrison & Henkel, 1970; Thompson, 1989c). A series of articles on the limits of statistical significance testing has even appeared on a seemingly periodic basis in recent editions of the American Psychologist (Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989; Rosenthal, 1991). The purposes of this paper are to elaborate three criticisms of overreliance on statistical significance testing, and to discuss three alternatives that may be useful to augment the evaluation of significance testing.

Three Criticisms of Statistical Significance Testing

Three of the various possible criticisms of conventional uses of statistical significance testing will be noted here. The first has generally not been explicated so directly, and the second two are essentially unrecognized by most researchers.

1. Statistical Significance Testing can be Tautological

Even some widely respected authors of prominent methodology textbooks at times take internally inconsistent positions with respect to the role that statistical significance testing should play in analysis (see book reviews by Thompson, 1987a, 1988d). And some dissertation authors may be disproportionately susceptible to excessive awe for significance tests (LaGaccia, 1991; Thompson, 1988b). But researchers who have had the experience of working with large samples (cf. Kaiser, 1976) soon realize that virtually all null hypotheses will be rejected at some sample size, since "the

null hypothesis of no difference is almost never exactly true in the population" (Thompson, 1987b, p. 14). As Meehl (1978, p. 822) notes, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Thus Hays (1981, p. 293) argues that "virtually any study can be made to show significant results if one uses enough subjects." Many researchers possess this insight[1], but somehow *do not integrate this knowledge into their paradigms* for conceptualizing or conducting research. Thus, the insight too rarely impacts actual practice.

A concrete heuristic example may serve to emphasize the impact that sample size can have on the outcomes of statistical significance tests. Presume that a researcher was working in a large school district, and analyzed data involving the district's 200,000 students. If the researcher decided to compare the mean IQ scores ($\overline{X}$ = 100.15, SD = 15) of 12,000 students located in one zip code with the mean IQ ($\overline{X}$ = 99.85, SD = 15) of the 188,000 remaining students residing in other zip codes, it would be decided that the two means differ to a statistically significant degree (Z_{CALC} = 2.12 > Z_{CRIT} = 1.96, $p<.05$). The less thoughtful researcher might suggest to school board members that special schools for gifted students should be erected in the zip code of the 12,000 students, since they are "significantly" brighter than their compatriots.

Alternatively, the more thoughtful researcher in such a situation would note that the standardized difference in these two means (.3/15 = 0.02) is trivial. The difference in the means (.3 =

one-third of one IQ point) is also substantially smaller than even just one standard error (used to construct a confidence interval capturing only 68% of the "true" scores, assuming measurement error is normally distributed) of an IQ measure with a reliability coefficient of 0.92, i.e., SEM = SD*((1-r)**.5) = 4.24. Such a thoughtful researcher would be reticent to extrapolate policy recommendations from every statistically significant result.

Although statistical significance is a function of at least seven interrelated features of a study (Schneider & Darcy, 1984), sample size is a basic influence on significance. To some extent significance tests evaluate the size of the researcher's sample--most researchers already know prior to conducting significance tests whether the sample in hand is large or small, so these outcomes do not always yield understanding that would be lost absent a significance test. As Thompson (in press-b) notes:

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. This tautology has created considerable damage as regards the cumulation of knowledge...

2. <u>Statistical Significance Testing can Invoke Somewhat Nonsensical Comparisons</u>

Researchers are frequently encouraged to employ statistical

significance tests in a linear or hierarchical sequence. For example, Keppel and Zedeck (1989) recommend that factorial ANOVAs should be conducted by testing and interpreting highest-order interaction effects, prior to evaluating main effects. But the different hypotheses in ANOVA can each involve different distributions of the sample size across different means, and consequently different power against Type II error. As Thompson (1991b, p. 503) notes:

> For example, in a 6 x 4 x 2 design with three subjects per cell, the omnibus three-way interaction involves 48 means each calculated over three subjects, while at the other extreme the C-way main effect involves two means each calculated over 72 subjects. Given differential power to detect various effects (which led to the recognition in the literature of Type IV error), the hierarchical approach guided exclusively by statistical tests conducted at a fixed alpha amounts to comparing apples and oranges.

3. <u>Sole Reliance on Statistical Significance Testing Creates Inescapable Dilemmas for Researchers</u>

Researchers who place an inordinate emphasis on statistical significance tests also often confront an inescapable dilemma, though most researchers do not recognize (or prefer to ignore) this dilemma. All statistical significance tests invoke certain assumptions. For example, ANOVA requires pooling the variances of the dependent variable across the cells of the design during the

calculation of the mean square used in the denominator of the fixed-effects F-test. This pooling is legitimate if and only if the variances of the dependent variable scores in all the cells are essentially equal. This is the well known "homogeneity (i.e., equality) of variance" assumption.

Similarly, as Thompson (in press-a) notes, ANCOVA is a three-stage analysis in which (a) regression weights for the covariate are derived *completely ignoring* group or cell membership of the subjects, (b) predicted dependent variable scores ($\hat{Y}$) are computed using the weights, and are then subtracted from the actual dependent variable scores (Y) of the subjects to yield an "e" score ("e_i" = $Y_i - \hat{Y}_i$) for each ith subject, and then (c) an ANOVA is conducted using the "e" scores as the dependent variable in place of the Y scores. As Loftin and Madison (1991) explain in some detail, this process is legitimate if and only if the regression equations for predicting Y with the covariate(s) are essentially the same, i.e., the "homogeneity of regression" assumption is met. Because a single regression equation, a single equation that is calculated completely ignoring group membership, is employed to statistically adjust the Y scores, this single equation can only reasonably be used if the equations for the different groups or cells are reasonably comparable, otherwise use of a "pooled" regression equation would be inappropriate.

Many researchers use statistical significance testing to evaluate *both* their preliminary methodological assumption hypotheses (e.g., the ANOVA homogeneity of variance assumption, the

ANCOVA homogeneity of regression assumption) and their substantive hypotheses (e.g., the mean dependent variable score of the treatment group equals that of the control group). These researchers hope to not reject the null hypotheses involving methodological assumptions (e.g., they want the dependent variable variances in the cells to all be equal), while they typically hope to reject their substantive hypotheses. But as Thompson (1991b, p. 504) notes, this creates a dilemma, since

> the same large sample size that yields power against Type II error in testing the substantive hypotheses of interest in ANCOVA [or ANOVA or the t-test] is also going to tend to yield statistically significant effects for the preliminary homogeneity of regression [or of variance] test.

Some researchers attempt to escape this dilemma by presuming that their methods are robust to the violation of their assumptions. This does not generally appear to be the case with respect to ANCOVA (Keppel & Zedeck, 1989). And the longstanding view that ANOVA was robust to the violation of the homogeneity of variance assumption has recently been called into some question, thanks to more sophisticated Monte Carlo studies conducted with more complicated designs, and with more simulation samples (e.g., Rogan & Keselman, 1977; Tomarkin & Serlin, 1986; Wilcox, Charlin & Thompson, 1986).

Three Alternatives to Supplement Statistical Significance Testing

None of this is to argue here that statistical significance

testing should be abandoned. It is useful to have some estimate, albeit a limited one, regarding the probability of a sample result, assuming that the sample came from a population in which the null was true.

But it is suggested that statistical significance testing has somewhat limited utility, and that greater attention should be focused on alternative analyses that are more central to the purposes of science, i.e., the accumulation of knowledge. Over the years various alternatives that might serve as substitutes for or augmentations of statistical significance tests have been proposed. For example, Serlin and Lapsley (1985) advocated placing an emphasis on confidence intervals, Bayesian approaches have been encouraged by others (e.g., Good, 1981), and somewhat less serious proposals have been presented by some (Salzman, 1989).

Three alternatives will be elaborated. Each is offered as an independent alternative to augment statistical significance testing, though all three could be used by a researcher conducting a given study. The first alternative has been discussed by various researchers, but is presented here in a more conceptual manner. The second alternative has been suggested in my previous work. The third alternative is more widely known by mathematical statisticians than by behavioral researchers.

1. Evaluating Result Importance by Consulting Effect Sizes

Statistical significance tests do not inform the researcher regarding the importance of results. Statistical significance tests evaluate the probability of an actual result, assuming that the

sample data come from a population in which the null hypothesis is exactly true. But an improbable result is not necessarily an important result, as Shaver (1985, p. 58) illustrates in his hypothetical dialogue between two teachers:

> Chris: ...I set the level of significance at .05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that *surely* must be important.
>
> Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?

Statistics can be employed to evaluate the probability of an event. But importance is a question of human values, and math cannot be employed as an escape (a la Rogers' Escape from Freedom) from the existential human responsibility for making value judgments. Like it or not, empirical science in inescapably a subjective business.

Many effect size estimates (e.g., Hays, 1981; Tatsuoka, 1973; Wherry, 1931) are available for researchers who wish to inform subjective judgment regarding result importance. The simplest effect sizes are analogous to the coefficient of determination (r^2). For example, in analysis of variance the sum of squares (SOS) for

an effect can be divided by the SOS total to compute the correlation ratio (also called eta squared). Such statistics inform the researcher regarding what proportion of variance in the dependent variable(s) is explained by a given predictor. The related effect size in regression is the R^2 statistic calculated by dividing the $SOS_{REGRESSION}$ by the SOS for the Y scores, i.e., SOS_{TOTAL}.

The simplest effect sizes are based on the data in hand and sample size and degrees of freedom are not considered as part of the calculations. However, all classical parametric methods are correlational (Knapp, 1978; Thompson, 1988a, 1991a) and do capitalize on sampling error as one part of their least squares analyses. This realization suggests that there are three major classes of effect size estimates: (a) biased overestimates, such as eta squared and R^2, (b) estimates that correct for positive bias in developing expectations for the likely effect size in the *population*, e.g., Hays' omega squared (see Maxwell, Camp & Arvey, 1981; Rosnow & Rosenthal, 1988), and (c) estimates employing corrections for the positive bias that also results when using least squares methods to estimate effect sizes likely to be realized in future *samples* from the population (Herzberg, 1969). From one perspective it might be argued (and has by some--see Stevens, 1986) that estimates in the last class are the most relevant, since in practice scientists extrapolate expectations from previous studies with samples and hope their results will be replicated in future studies with samples.

Positive bias, and consequently the related statistical

corrections for bias, both tend to be larger as either effect sizes or sample sizes (especially relative to the number of variables) become smaller, as illustrated by Thompson (1990). Thus, with a very large effect size or a large sample size, it will matter less which, if any, corrections the researcher applies in estimating effect sizes (cf. Carter, 1979; Pedhazur, 1982, p. 148).

Indeed, for various sample sizes ranging from 10 to 320, for uncorrected R^2 effect sizes ranging from 1% to 90%, and for numbers of predictor variables ranging from one to 10, uncorrected R^2 values and values corrected for shrinkage in estimating the *population* effect size (e.g., Olkin & Pratt, 1958; Wherry, 1931) have a product-moment correlation of about .90, while uncorrected R^2 values and values corrected for shrinkage in estimating future *sample* effect sizes (e.g., Herzberg, 1969; Lord, 1950) have a product-moment correlation of about .98. With respect to their sizes, though the estimates tend to be very highly correlated across designs, for a given design the uncorrected estimate is always largest, while the corrected estimate of future *sample* values tends to be smallest (Fisk, 1991).

Rosenthal's (1991, p. 1086) statement about one collection of effect size estimates is true beyond even the context of his discussion: "There is no right answer to the question of which of these indices is best or most useful under all conditions." All effect size estimates have some limits, and like all statistics must be interpreted reflectively (McGraw, 1991).

The biggest objections to using effect sizes occur with

respect to the use of effect sizes in certain applications in the ANOVA family. The case for interpreting effect sizes via r^2 analogs is that (a) these effect sizes are in a metric that allows direct comparison of values across different hypotheses or different studies (unlike r, for example, where r = 1.0 is not twice the size of r = 0.5) and (b) they indicate how much of the area (measured in squared units, just like carpet or floor tile) of the dependent variable in univariate studies, expressed again in squared units as the sum-of-squares (SOS), is explained by a given combination of independent or predictor variables.

However, in ANOVA applications these effect sizes are most useful when the levels in the ways or factors are (a) all the levels that make up the ways (e.g., male vs female for the way, gender) or (b) a random or representative selection of the levels in the full universe of levels that defines given ways (e.g., 5 vs 8 vs 33 minutes of computer instruction randomly selected from all possible intervention times available in a 55 minute class period). The latter case represents what are termed "random effects" (Glass & Hopkins, 1984, Chapter 19), and though some researchers are only familiar with "fixed effects" models, the use of random effects models in ANOVA work has been strongly advocated by some researchers (Clark, 1973, 1976; Wike & Church, 1976).

Some researchers (e.g., Glass & Hakstian, 1969) suggest that effect sizes such as omega squared are not useful unless the levels in ANOVA ways are all the possible levels of ways or are a representative sample of them. However, even in other cases I

believe effect size estimates are useful, as long as one remembers that (a) these effect sizes tell the researcher how much dependent variable variance (SOS) the differences in a given collection of levels explain, but that (b) these effects might change if a different collection of levels was used. The fact that a statistic has some limits under certain circumstances does not mean that the statistic must be completely abandoned, especially since all statistics have limits.

Cohen's (1988) perusal of published research suggests that a correlation ratio of around 25% (r=.5) should be considered large in terms of typical findings across disciplines. The empirical meta-analytic work of Glass (1979, p. 13) and others (Olejnik, 1984, p. 43) has also led to similar conclusions regarding typical effect sizes. Although it is sometimes useful to know what effect sizes are typical in social science generally and in certain areas of inquiry more particularly, the importance of an effect size ultimately depends upon the particular context of a specific study, and on an individual researcher's personal value system, rather than on typicality. For example, an effect size of 3% in an intervention study involving a vaccine for AIDS would be deemed valuable by researchers (a) who value human life greatly, and (b) who believe that most AIDS intervention studies have to date yielded 0% effect sizes, even though (c) interventions in social science generally may yield effects as large as 25%.

2. Evaluating Results in a Sample Size Context

A second strategy for augmenting interpretation of statistical

significance tests involves evaluating significance test results in a sample size context. The researcher can estimate roughly at what smaller sample size a statistically significant fixed effect size would no longer be significant, or conversely, at what larger sample size a nonsignificant result would become statistically significant (Thompson, 1989a).

Table 1 illustrates this application. The table presents significance tests associated with varying sample sizes and what are large (33.0%) effect sizes at least with respect to their typicality (Cohen, 1988; Glass, 1979; Olejnik, 1984). The table can be viewed as presenting results for either a multiple regression analysis involving two predictor variables (in which case the "r^2" effect size would be called the squared multiple correlation coefficient, R^2) or an analysis of variance involving an omnibus test of differences in three means in a one-way design (in which case the "r^2" effect size would be called the correlation ratio or eta squared).

INSERT TABLE 1 ABOUT HERE.

The table presents results for fixed effect sizes but increasing sample sizes (4, 13, 23, or 33). For the 33.0% effect size reported in Table 1, the result becomes statistically significant when there are somewhere between 13 and 23 subjects in the analysis.

The researcher who does not genuinely understand statistical significance would differentially interpret the effect size of

33.0% when there were 13 versus 23 subjects in the analysis. Yet the effect sizes within the table are fixed. Empirical studies of scholarly practice indicate that superficial understanding of significance testing has actually led to serious distortions such as researchers interpreting statistically significant results involving small effect sizes while ignoring nonsignificant results involving large effect sizes (Craig, Eison & Metze, 1976)!

Since sample effect sizes are positively biased partly as a function of sample size (with more bias in smaller samples), a more elegant approach would invoke corrections for the effect size estimates for the various sample sizes as part of this logic. Statistically simpler corrections (e.g., Wherry, 1931) might be employed, or more accurate but more computationally complicated corrections (e.g., Browne, 1975) might be used. Cattin (1980), Mitchell and Klimoski (1986), and Schmitt (1982) review some of the choices.

However, the purpose of this approach is not to identify the exact results that would occur with a different sample size, assuming exactly the same effect size. Rather, the approach focuses on establishing a general ballpark for interpreting statistical significance tests in a sample size context. Thus, the analysis should not be overinterpreted, any more than the results of conventional statistical significance testing should be overinterpreted.

3. Interpreting Results Based on Likelihood of Replication

A third strategy emphasizes interpretation based on the

estimated likelihood that results will replicate. This emphasis is compatible with the basic purpose of science: isolating conclusions that replicate under stated conditions. Notwithstanding some misconceptions to the contrary, statistical significance tests do not evaluate the probability that results will generalize (Carver, 1978).

The ultimate test of replicability is an actual replication study, but it is not always convenient to conduct a replication prior to interpreting results. Three general strategies provide the next best evaluation of replicability. But since all three strategies are typically based on a single sample of subjects in which the subjects usually have much in common (e.g., point in time of measurement, geographic origin) relative to what they would have in common with a separate sample, the three methods all yield somewhat inflated estimates of replicability. Because inflated estimates of replicability provide a better estimate of replicability than no estimate at all (i.e., statistical significance testing), these procedures can still be useful in focusing on the *sine qua non* of science.

The first two methods both involve splitting the sample into two or more subsamples, and then <u>empirically</u> comparing results across sample splits. The first method is *cross-validation*, and typically involves splitting the sample into two roughly equally-sized groups (Huck, Cormier & Bounds, 1974, pp. 159-160). Thompson (1989c) provides step-by-step illustrations of this approach.

The second approach invokes the *jackknife* methods elaborated

by Tukey and his colleagues (cf. Crask & Perreault, 1977). This approach involves conducting separate analyses with various groups of subjects each deleted from the analysis (usually each $n_{DROPPED}$ = 1) one time and conducting all possible k analyses under these conditions (if $n_{DROPPED}$ = 1, k = n) in addition to an analysis in which no subjects are dropped. Daniel (1989) provides a tutorial on this method.

But a particularly powerful strategy for evaluating result replicability invokes the *bootstrap* methods developed by Efron and his colleagues (cf. Diaconis & Efron, 1983; Efron, 1979; Lunneborg, 1990). Conceptually, these methods involve copying the data set over again and again many many times into an infinitely large "mega" data set. Then hundreds or thousands of different samples are drawn from the "mega" file, and results are computed separately for each sample and then averaged.

The method is powerful because the analysis considers so many configurations of subjects (including configurations in which a subject may be represented several times or not at all) and informs the researcher regarding the extent to which results generalize across different types of subjects. Lunneborg (1987) has offered some excellent computer programs that automate this logic for univariate applications; Thompson (1988c) provides similar software for multivariate applications. Recently, user-friendly PC bootstrap software has become available from publishers around the world.[2]

Table 2 presents a small data set that can be used to illustrate both conventional and bootstrap estimation. The table

presents Z score values from each of 11 subjects on two variables. The data set is unrealistically small for actual bootstrap applications, but has heuristic value and is sufficiently manageable in size to allow interested readers to replicate the results reported here.

INSERT TABLE 2 ABOUT HERE.

All statistical tests invoke four estimates. The first is a *single* statistic estimating a single population parameter calculated from the sample data in hand. The remaining three estimates are calculated not from the data in hand, but rather from entirely different data (called the sampling distribution) conceptually involving *multiple* repeated samplings of the parameter estimate from a population. These four estimates are: (a) the single parameter estimate (e.g., $\bar{X}$, r) derived from a sample believed to be representative of a population; (b) the second moment about the mean of multiple estimates of the parameter of interest (i.e., the standard deviation (SD) of the repeated sampled estimates, called the standard error (SE_E) of the estimated statistic); (c) the third moment about the mean of multiple estimates of the parameter (i.e., the coefficient of $skewness_E$); and (d) the fourth moment about the mean of multiple estimates of the parameter (i.e., the coefficient of $kurtosis_E$).

Many researchers recognize the use of the first two statistics in their statistical analyses. Thus, researchers using LISREL and EQS analyses routinely pay more attention to parameter estimates

that are greater than the individual standard errors of given estimates. As Kerlinger (1986, chapter 12) explains in some detail, test statistics also invoke the ratio of a parameter estimate to the SE_E. For example, researchers often use a t-test to evaluate the null hypothesis that a mean equals zero. For a sample of size n, the SD of infinitely many samples of size n from a population in which the mean is zero (i.e., $SE_{\bar{X}}$) would be approximately $SD_X/(n^{**}.5)$. The test statistic, $t_{CALCULATED}$, for this research situation is calculated as the ratio of $\bar{X}$ / $(SD_X/(n^{**}.5))$.

The use of the third and fourth statistics is not so explicit. But when we evaluate the probability of our sample result, $p_{CALCULATED}$ given an assumption that the null is true, we usually compare our result against the α (or the $\alpha/2$) percentile of the test statistic, and the skewness and the kurtosis of this sampling distribution are part of what dictates what will be the value the α%ile of the test distribution. Of course, conventional confidence intervals employ exactly the same elements as statistical significance testing, and do make the use of all four estimates explicitly obvious (Glass & Hopkins, 1984, section 11.7).

Table 3 presents the calculated sample statistic, r = +.560, for the Table 2 data, and this same result expressed using Fisher's r-to-Z transform. Table 3 also presents calculated SE_{Z_r} (.354), calculated *assuming* both that population value of Z_r is zero, and that the sampling error is normally distributed (skewness and kurtosis both equal 0) about Z_r. Given these assumptions[3], we can infer that roughly 95% of the samples from the population will fall

between Z_r = -.061 (r = -.060) and Z_r = +1.325 (r = +.868).

INSERT TABLE 3 ABOUT HERE.

Because the 95% confidence interval subsumes 0, the r of +.560 is not deemed statistically significant.[4] The use of the 97.5%ile from the Z distribution, 1.96, in the confidence interval calculations is where both skewness and kurtosis coefficients are invoked, for the 97.5%ile of Z scores will be 1.96 only if skewness and kurtosis are both zero.

However, *it is contradictory to be willing to use the sample to derive our (a) parameter estimate, and to be unwilling to let the sample offer similar insight regarding the (b) SE of our estimate, and regarding the (c) skewness and (d) kurtosis of sampled estimates*. One way explore our data regarding the latter three estimates is to conduct a bootstrap analysis, i.e., we momentarily treat our sample data as if it constituted the population and we draw numerous (usually at least a thousand) random samples from the sample to infer what the sampling distribution looks like. To mimic randomly sampling our data with n subjects from the population, we do all our "resampling" from our mock population by drawing random samples with replacement from our data in hand, and to honor our research situation each resample is drawn to also have exactly size n.

The Table 2 data can be used to illustrate this application and its potential benefits. These estimates were developed using the software available from Lunneborg (1987), and were based on

1,000 samples with replacement. As reported in Table 3, the standard deviation of the 1,000 estimates of r was .173--this is the *empirical* estimate of SE_r, and is considerably smaller than the estimate of the SE (SE_{Z_r} = .354, SE_r = .339) derived based on assumptions. The bootstrap results were also useful in alerting the researcher to the fact that the sampling distribution may not be normal, e.g., the distribution may be negatively skewed.

The bootstrap approach can be employed to yield a variety of confidence intervals, which vary as a function of the assumptions they make about the sampling distribution. The three estimates calculated by the Lunneborg (1987) program for the Table 2 data are reported in Table 3. The "bias corrected" estimate makes the fewest assumptions regarding the sampling distribution (Lunneborg, 1987, p. 54), that is, relies most upon the empirical findings from resampling. Since none of the confidence intervals subsume zero, the bootstrap results employing an empirically estimated sampling distribution, unlike the conventional approach, yields a statistically significant result.

Of course, bootstrap and other methods that focus on the invariance or the generalizability of results are no more magical than is classical statistical significance testing itself. No analytic methods can magically take us beyond the limits of our data. We use methods to explore data in various ways, not to make data more than they can be.

Discussion

The conflict between the quantitative and qualitative research

paradigms (Thompson, 1989b) has helped researchers in both schools recognize and acknowledge that the researcher is inherently "caught up in the web of circumstances under study; he [sic] cannot escape his role as an actor in society" (Piel, 1978, p. 9). Researchers ought to abandon any illusion "that it is adherence to a series of established procedures which prevent the self from disrupting or distorting this 'journey of the facts'" (Smith, 1983, p. 10). But moving some researchers in this direction may be a difficult proposition, given

> that one of the hardest tasks statisticians face is persuading investigators to look at their data. This is a situation that is not likely to soften, given the epidemic rise in the number of *p*-value software statisticians. (Bartko, 1991, p. 1089)

More researchers need to recognize the limits of statistical significance tests, and ought to augment these analyses.

Resistance to relying less on the p values from statistical significance testing cannot be successfully rationalized on the grounds that significance tests yield any payoff in objectivity. As Berger and Berry (1988) argue, such a view would be an "illusion", since "objectivity is not generally possible in statistics" (p. 165). Huberty and Morris (1988, p. 573) concur, noting that "As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness!"

The *single* study is inherently governed by subjective passion (Kerlinger, 1986, p. vii), and by ideology even as regards analytic

choice (Cliff, 1987, p. 349). The protection for fledgling efforts to obtain insight does not arise from lockstep adherence to a flowchart sequence of design and analytic choices. Scientific progress is grounded on impassioned observation (Thompson, 1989b, p. 37). The protection against the potentially negative consequences of these passions occur not from feigned objectivity, but arise in the *aggregate* across studies from an emphasis on replication (Neale & Liebert, 1986, p. 290).

It has not been said here that statistical significance testing should be abandoned. Rather, it has been suggested that statistical tests should not be overinterpreted, and that these tests can be usefully augmented by analyses that bear more directly upon the cumulation of knowledge.

Otherwise, obsession with statistical significance will continue to lead to editorial practices favoring articles that report statistically significant outcomes (Rosenthal, 1979). This is comforting in that it creates a bias against reports of Type II errors, since by definition statistically significant results cannot represent Type II errors (Thompson, 1987b). But, in the context of this bias, the greater likelihood of reporting statistically significant results that are in fact Type I errors is problematic, "because investigators generally cannot get their failures to replicate published, [and so] Type I errors, once made, are very difficult to correct" (Clark, 1976, p. 258).

Researchers who fail to obtain statistically significant results may abandon lines of inquiry (Greenwald, 1975), perhaps

even when such results are artifacts of insufficient power against Type II error. Researchers who fail to obtain statistically significant results may also decline to submit reports for publication (Rosenthal, 1979). Finally, even when such reports are submitted for review, such reports tend to be unfavorably received (Atkinson, Furlong & Wampold, 1982).

The adherence to worship at the temple of statistical significance testing, described vividly by Rosnow and Rosenthal (1989), cannot be defended on grounds of either tradition or an unwillingness to admit the error of past ways. Social science is a subjective business, and no analytic method can make it otherwise. There are several anaytic strategies that can be usefully employed to augment the results of statistical significance testing, and these methods may be more relevant to efforts to cumulate knowledge.

References

Atkinson, D.R., Furlong, M.J., & Wampold, B.E. (1982). Statistical significance, reviewer evaluations, and scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

Bartko, J.J. (1991). Proving the null hypothesis. American Psychologist, 46, 1089.

Berger, J.O., & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. American Scientist, 76, 159-165.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Carter, D.S. (1979). Comparison of different shrinkage formulas in estimating population multiple correlation coefficients. Educational and Psychological Measurement, 39, 261-266.

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cattin, P. (1980). Note on the estimation of the squared cross-validated multiple correlation of a regression model. Psychological Bulletin, 87, 63-65.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12, 335-359.

Clark, H.H. (1976). Reply to Wike and Church. Journal of Verbal Learning and Verbal Behavior, 15, 257-261.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt

Brace Jovanovich.

Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and omega-squared. Bulletin of the Psychonomic Society, 7, 280-282.

Crask, M.R., & Perreault, W.D., Jr. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Daniel, L.G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 305 382)

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7, 1-26.

Fisk, Y.H. (1991, April). Various approaches to effect size estimation. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED forthcoming)

Glass, G.V. (1979). Policy for the unpredictable (uncertainty

research and policy). Educational Researcher, 8(9), 12-14.

Glass, G.V, & Hakstian, A.R. (1969). Measures of association in comparative experiments: Their development and interpretation. American Educational Research Journal, 6, 403-414.

Glass, G.V, & Hopkins, K.D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Good, I.J. (1981). Some logic and history of hypothesis testing. In J. Pitt (Ed.,), Philosophy in economics (pp. 149-174). Dordrecht, Holland: Reidel.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Herzberg, P.A. (1969). The parameters of cross validation. Psychometrika, Monograph supplement, No. 16.

Huberty, C.J. (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Huberty, C.J, & Morris, J.D. (1988). A single contrast test procedure. Educational and Psychological Measurement, 48, 567-578.

Huck, S. W., Cormier, W. H., & Bounds, Jr., W. G. (1974). Reading statistics and research. New York: Harper & Row.

Kaiser, H.F. (1976). [Review of Factor analysis as a statistical method]. Educational and Psychological Measurement], 36, 586-589.

Keppel, G., & Zedeck, S. (1989). Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches. New York: W.H. Freeman.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart and Winston.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

LaGaccia, S.S. (1991). Methodology choices in a cohort of education dissertations. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 149-158). Greenwich, CT: JAI Press.

Loftin, L.B., & Madison, S.Q. (1991). The extreme dangers of covariance corrections. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 133-147). Greenwich, CT: JAI Press.

Lord, F. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample (Research Bulletin 50-110). Princeton, NJ: Educational Testing Service.

Lunneborg, C.E. (1987). Bootstrap applications for the behavioral sciences. Seattle: University of Washington.

Lunneborg, C.E. (1990). [Review of Computer intensive methods for testing hypotheses]. Educational and Psychological Measurement, 50, 441-445.

Maxwell, S.E., Camp, C.J., & Arvey, R.D. (1981). Measures of strength of association: A comparative examination. Journal of Applied Psychology, 66, 525-534.

McGraw, K.O. (1991). Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?". American Psychologist, 46, 1084-1086.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Mitchell, T. W., & Klimoski, R. J. (1986). Estimating the validity of cross-validity estimation. Journal of Applied Psychology, 71, 311-317.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Neale, J.M., & Liebert, R.M. (1986). Science and behavior: An introduction to methods of research (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Olejnik, S.F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Olkin, I., & Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. Annals of Mathematical Statistics, 29, 201-211.

Pedhazur, E.J. (1982). Multiple regression in behavioral research (2nd ed.). New York: Holt, Rinehart and Winston.

Piel, G. (1978). Research for action. Educational Researcher, 7(2),

8-12.

Rogan, J.C., & Keselman, H.J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. American Educational Research Association, 14, 493-498.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. Journal of Counseling Psychology, 35, 203-208.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Salzman, K.L. (1989). A significantly significant approach to significant research findings: The Salzman All-Significant F test. In G.C. Ellenbogen (Ed.), The primal whimper (pp. 158-162). New York, NY: Guilford Press.

Schmitt, N. W. (1982, August). Formula estimation of cross-validated multiple correlation. Paper presented at the annual meeting of the American Psychological Association, Washington, DC. (ERIC Document Reproduction Service No. ED 227 137)

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation

Review, 8, 573-582.

Serlin, R.C., & Lapsley, D. (1985). Rationality in psychological research: The good-enough principle. American Psychologist, 40, 73-83.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Smith, J.K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. Educational Researcher, 12(3), 6-13.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Erlbaum.

Tatsuoka, M.M. (1973). An examination of the statistical properties of a multivariate measure of strength of relationships. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Thompson, B. (1987a). [Review of Foundations of behavioral research (3rd ed.)]. Educational Research and Measurement, 47, 1175-1181.

Thompson, B. (1987b, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Education Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1988a, April). Canonical correlation analysis: An explanation with comments on correct practice. Paper presented at the annual meeting of the American Educational Research

Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B. (1988b, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1988c). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. Educational and Psychological Measurement, 48, 681-686.

Thompson, B. (1988d). [Review of Analyzing multivariate data]. Educational and Psychological Measurement, 48, 1129-1135.

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). The place of qualitative methods in contemporary social science: The importance of post-paradigmatic thought. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 1-42). Greenwich, CT: JAI Press.

Thompson, B. (1989c). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1990). Finding a correction for the sampling error in multivariate measures of relationship: A Monte Carlo study. Educational and Psychological Measurement, 50, 15-31.

Thompson, B. (1991a). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1991b). [Review of Data analysis for research designs]. Educational and Psychological Measurement, 51, 500-510.

Thompson, B. (in press-a). Misuse of ANCOVA and related "statistical control" procedures. Reading Psychology.

Thompson, B. (in press-b). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development.

Tomarkin, A.J., & Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 88, 90-99.

Wherry, R.J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 440-451.

Wike, E.L., & Church, J.D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy". Journal of Verbal Learning and Verbal Behavior, 15, 249-255.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F' statistics. Communications and Statistics, 15, 933-943.

Footnotes

[1]Though many researchers possess this elementary insight, not all do. A blind referee for a respected journal reviewing a manuscript making this point noted, "It certainly is <u>not</u> the case, as the au. contends, that with a huge enough *n*, the null hypothesis will inevitably be rejected; what about,[sic] psychokinesis, Mendelian hypotheses for progeny, superstitions?" However, even if we admit only an infinitesimal measurement error influence that creates a difference in two means at the 39th place to the right of the decimal in a population in which the two means are exactly equal (or a population with no measurement error at all in which two means differ only at the 39th decimal place), large enough samples from the population will detect these differences as being statistically significant.

[2]Examples of such software and the distributors of the software include: (a) "Resampling Stats", distributed by Resampling Stats, 612 N. Jackson, Arlington, VA 22201; (b) "Statistical Calculator", distributed by Erlbaum, 27 Palmeira Mansions, Church Road, Hove East Sussex BN3 2FA, United Kingdom; (c) SPIDA, distributed on behalf of its Australian author by SERC, 1107 NE 45th--Suite 520, Seattle, WA 98105; and (d) the menu-driven program, BOJA, distributed by iecProGAMMA, P.O. Box 841, 9700 AV Groningen, The Netherlands.

[3]It is actually contradictory to calculate SE_{Z_r} based on an assumption that $\underline{r} = 0$, and to then use SE_{Z_r} to calculate confidence intervals for $\underline{r} \neq 0$, unless one only wishes to test H_0: $\underline{r} = 0$. In

this case conceptually the CI is really being constructed around 0 (and not r), and the test is whether the point estimate, r, falls within the interval. However, in practice we usually consider this estimation procedure to be "close enough".

[4]Nor is the result statistically significant when it is evaluated using the more powerful two-tailed t test.

Table 1
Statistical Significance at Various Sample Sizes
for a Fixed Effect Size

Source	SOS	r^2	df	MS	Fcalc	Fcrit	Decision
SOSexplained	**331.2**	**33.0%**	**2**	**165.600**	0.246	200.00	Not Rej
SOSunexplained	**672.3**		1	672.300			
SOStotal	**1003.5**		3	334.500			
SOSexplained	**331.2**	**33.0%**	**2**	**165.600**	2.463	4.10	Not Rej
SOSunexplained	**672.3**		10	67.230			
SOStotal	**1003.5**		12	83.625			
SOSexplained	**331.2**	**33.0%**	**2**	**165.600**	4.926	3.49	Rej
SOSunexplained	**672.3**		20	33.615			
SOStotal	**1003.5**		22	45.614			
SOSexplained	**331.2**	**33.0%**	**2**	**165.600**	7.390	3.32	Rej
SOSunexplained	**672.3**		30	22.410			
SOStotal	**1003.5**		32	31.359			

Note. As sample size increases, tabled "critical F" values get smaller. Additionally, as sample size increases, error df gets larger, mean square error gets smaller, and thus "calculated F" also gets larger. Entries in **bold** remain fixed for the purposes of these analyses.

Table 2
Hypothetical Data Used to Illustrate Bootstrap
Evaluation of an Estimate of r

ID	Y	X
1	.18	.20
2	.54	1.88
3	-.49	-.76
4	.92	.42
5	.22	.32
6	.75	-.56
7	.66	1.55
8	-2.65	-1.21
9	-.51	-.66
10	.47	-.96
11	-.09	-.21
r_{YX}	.560	
Z_r	.632	

Note. $Z_r = 1.1513\ (\ln\ ((1 + |r|) / (1 - |r|)))$

$1.1513\ (\ln\ ((1 + .560) / (1 - .560)))$

$1.1513\ (\ln\ (\ 1.560 / .440))$

$1.1513\ (\ln\ (\ 3.541))$

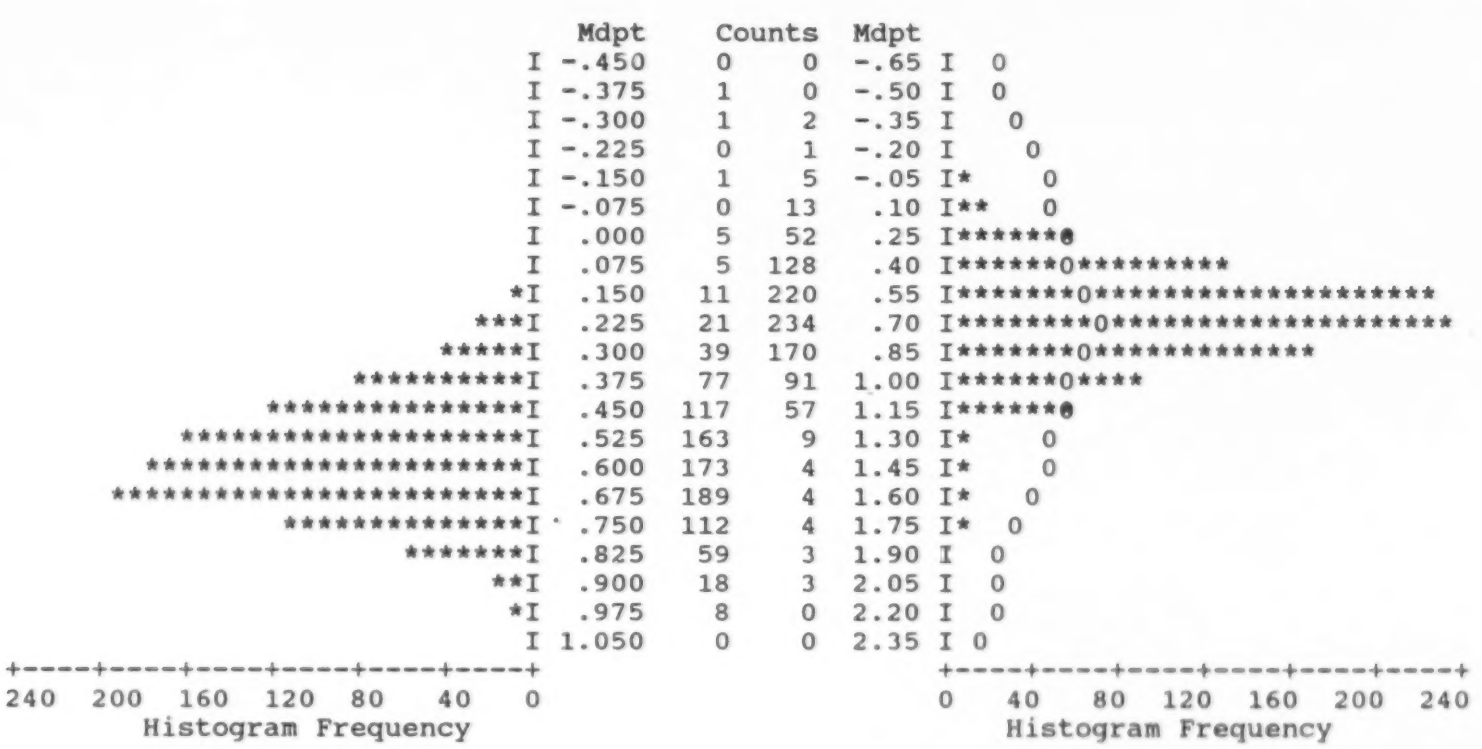
$1.1513\ (.549) = .632$

Table 3
Conventional and Bootstrap Significance Tests
for $\underline{r}$=.560 for the Table 2 Data

Sampling Statistics/ Significance Tests	Classical Estimates Based on *Statistical Assumptions*	*Empirically* Based Bootstrap Estimates
Second Moment of the Sampling Distribution		
SE_{Z_r}	.354[a]	
SE_r	.339[b]	.173
Third Moment of the Sampling Distribution		
Coefficient of Skewness of $\underline{r}$	.000 (*assumed*)	-.780
Third Moment of the Sampling Distribution		
Coefficient of Kurtosis of $\underline{r}$	.000 (*assumed*)	1.895
Density of the Sampling Distribution		
90.0%ile of Z_r	1.282 (*assumed*)	1.037
95.0%ile of Z_r	1.645 (*assumed*)	1.164
97.5%ile of Z_r	1.960 (*assumed*)	1.324
95% Confidence Intervals		
About Z_r	-.061 to 1.325[c]	
About $\underline{r}$	-.060 to 0.868[d]	+.220 to +.899[e] +.188 to +.868[f] +.082 to +.822[g]

[a]*Calculated* as SE_{Z_r} = 1 / ((n - 3) ** .5) = 1 / ((11 - 3) ** .5) =
1 / (8 ** .5) = 1 / 2.828 = .354.
[b]*Calculated* as SE_{Z_r} = .354 converted back into SE_r = .339.
[c]*Calculated as* $CI_{95\%}$ about $Z_r = Z_r - (1.960 * SE_{Z_r})$ to $Z_r + (1.960 * SE_{Z_r})$
= .632 - (1.960 * .354) to .632 + (1.960 * .354)
= .632 - .693 to .632 + .693
[d]The conversion of $\underline{r}$ expressed as Fisher's Z transform back into $\underline{r}$.
[e]$CI_{95\%}$ calculated using symmetric or normal theory approach.
[f]$CI_{95\%}$ calculated using percentile method.
[g]$CI_{95\%}$ calculated using bias corrected method.

Figure 1
Bootstrap Estimates of $\underline{r}$ Based on 1,000 Random Resamplings



Note. Each asterisk represents approximately eight cases. The distribution of 1,000 bootstrap estimates of $\underline{r}$ is presented to the left, while the distribution of the Fisher's Z transformation of these 1,000 estimates is presented to the right. The normal distribution of samples of Z_r, expected given the classical statistical assumptions that sampling error is distributed normally about the estimate, is also presented in the histogram on the right.

41

42